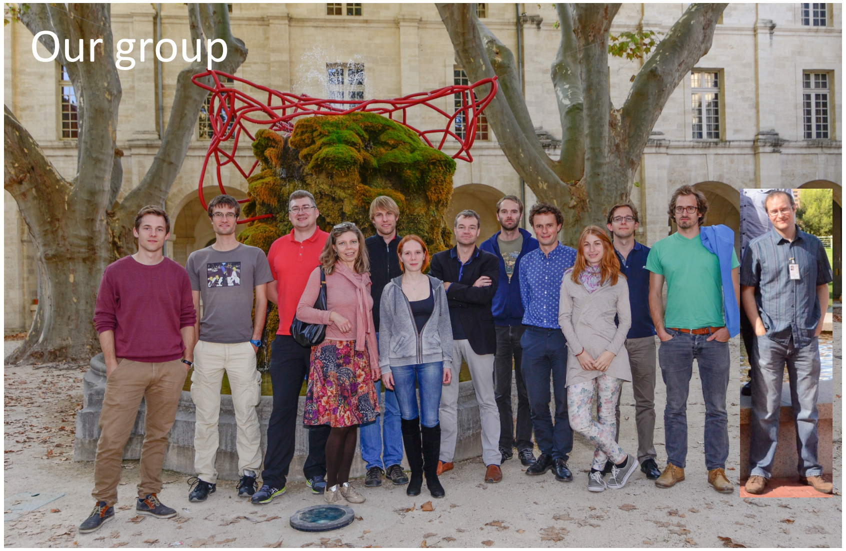
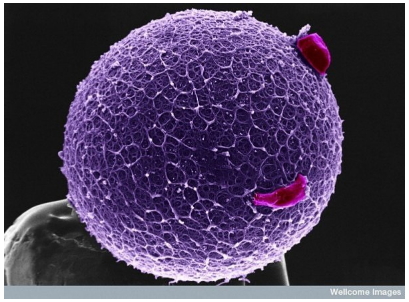
Tools for modelling regulatory genomics data in terms of predicted regulatory sites on the DNA

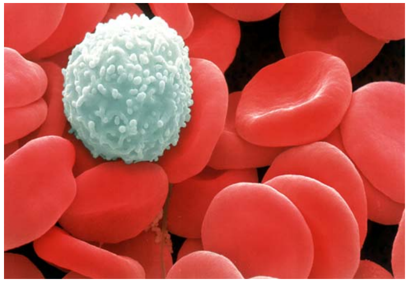


Erik van Nimwegen
Biozentrum, University of Basel,
and Swiss Institute of Bioinformatics

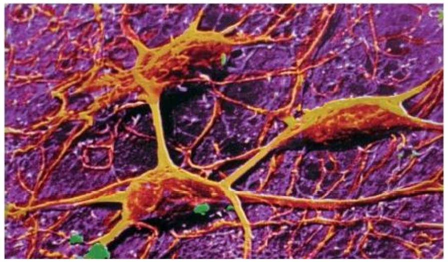
How is the regulatory code in the DNA `read out’ to control cell fate and identity?



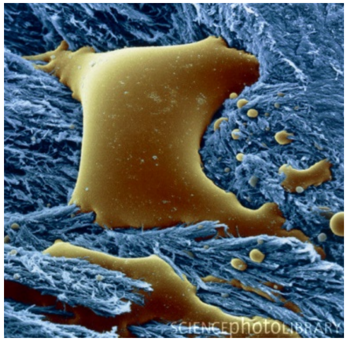
egg cell with 2 coronal cells



white and red blood cells



three neurons



osteoclasts

How do gene regulatory networks function as *systems*.

- What is a cell type?
- How is cell identity stabilized?
- Where is the information? What does not matter?

My worries

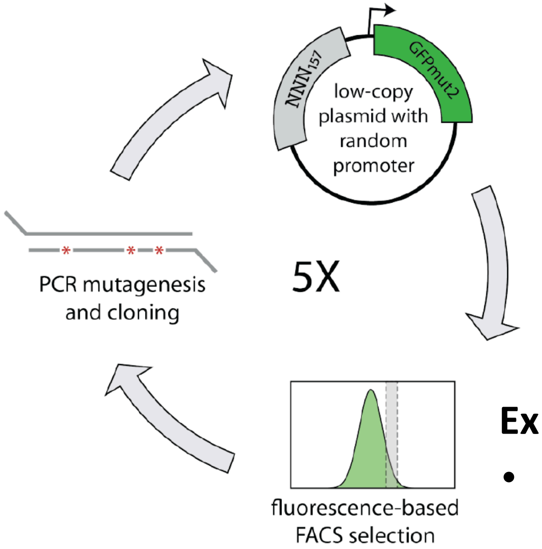
- We think we know/measure a lot, but there is orders of magnitude more we do not know.
- High-throughput measurements full of artefacts and biases that we poorly understand.
- Nowhere near the ability to meaningfully model what is going on.

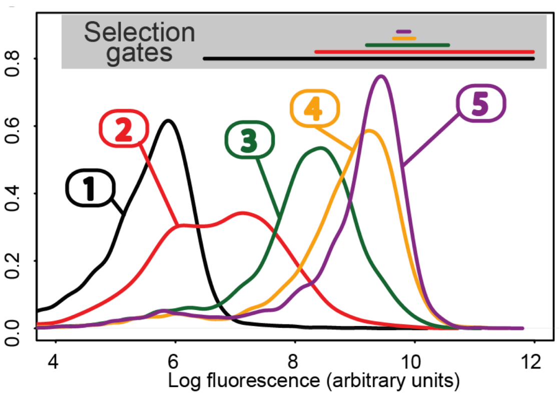
What useful things can a serious computational biologist do?

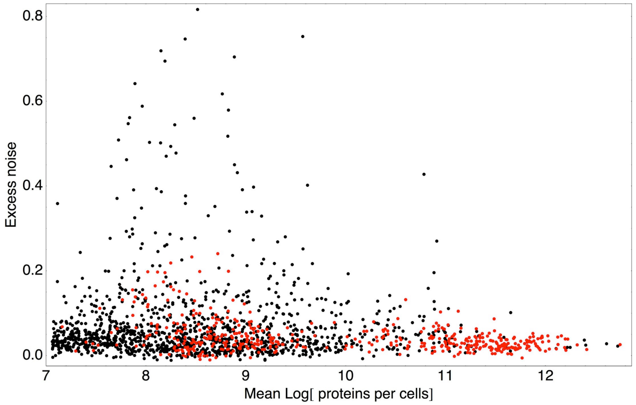
BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Expression noise facilitates the *de novo* evolution of gene regulation


NNN157
GFPmut2
low-copy plasmid with random promoter
PCR mutagenesis and cloning
5X
fluorescence-based FACS selection



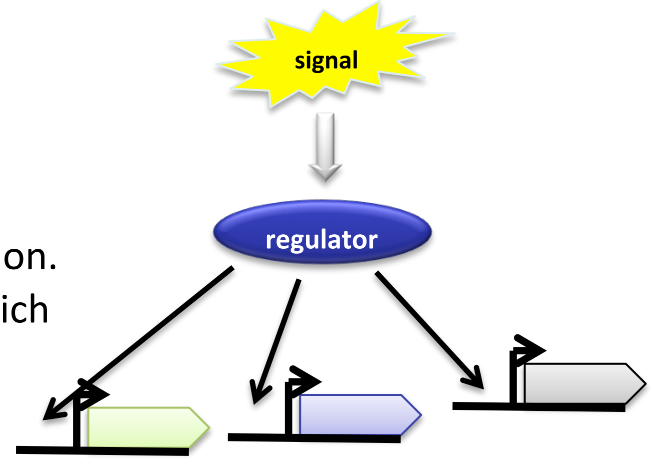


Experimental observations

- We evolved synthetic promoters *de novo* in *E. coli* under carefully-controlled selective conditions.
- No evidence *E. coli* promoters have been selected to lower noise.
- Promoters of regulated genes have been selected to *increase* noise.

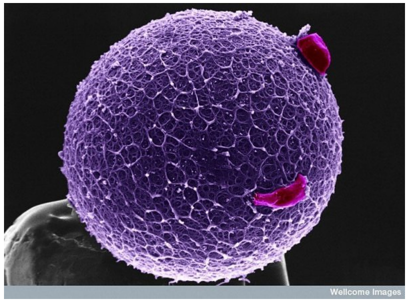
Theory

- Coupling a regulator to a target promoter has two effects:
  1. Condition-response.
  2. Noise-propagation.
- Noise-propagation alone can act as a rudimentary form of regulation.
- Accurate regulation can evolve smoothly along a continuum in which noise-propagation and condition-response act in concert.
- Explains the general association between noise and regulation.

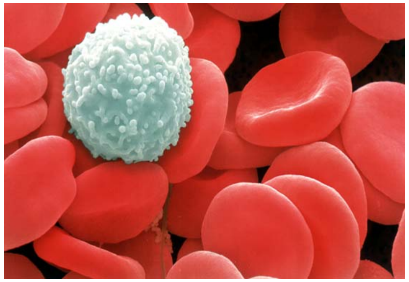


Wolf, Silander, van Nimwegen, *eLife*, 2015

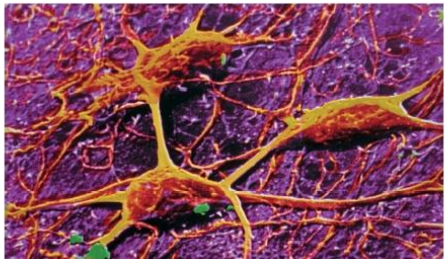
How is the regulatory code in the DNA `read out’ to control cell fate and identity?


Wellcome Images

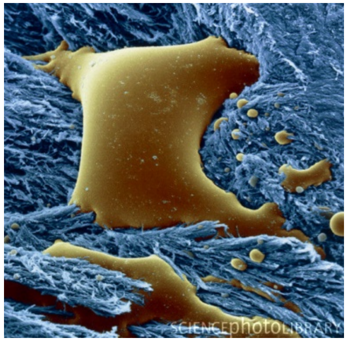
egg cell with 2 coronal cells



white and red blood cells



three neurons


SCIENCEphotoLIBRARY

osteoclasts

How do gene regulatory networks function as *systems*.

- What is a cell type?
- How is cell identity stabilized?
- Where is the information? What does not matter?

My worries

- We think we know/measure a lot, but there is orders of magnitude more we do not know.
- High-throughput measurements full of artefacts and biases that we poorly understand.
- Nowhere near the ability to meaningfully model what is going on.

What useful things can a serious computational biologist do?

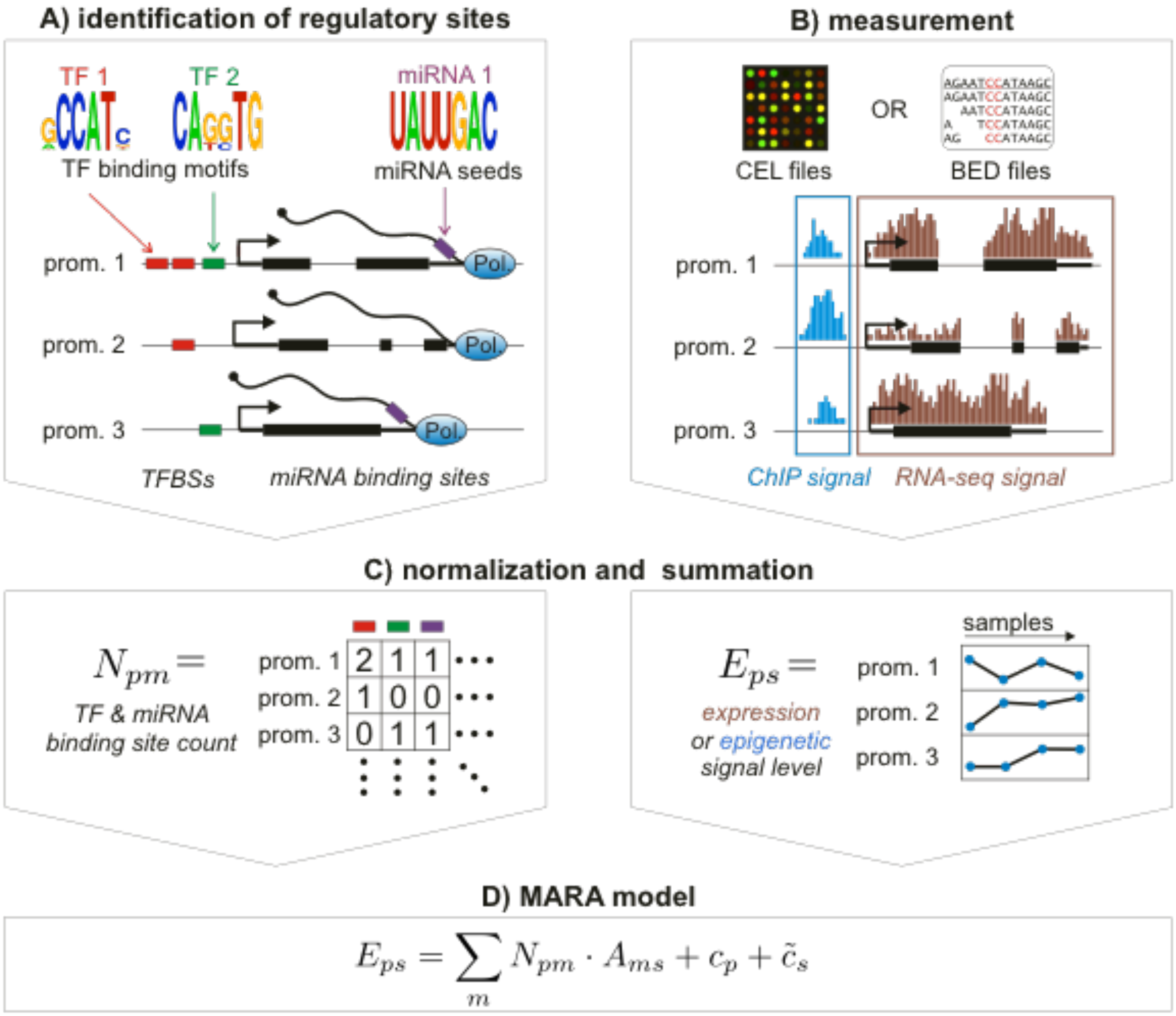
Develop simple, robust, and transparent methods that help guide experimental efforts.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Motif Activity Response Analysis:

Modeling gene expression and chromatin state in terms of TFBS using a linear model



prom. 1	2	1	1	⋯
prom. 2	1	0	0	⋯
prom. 3	0	1	1	⋯
	⋮	⋮	⋮	⋱

$$E_{ps} = \sum_{m} N_{pm} \cdot A_{ms} + c_p + \tilde{c}_s$$

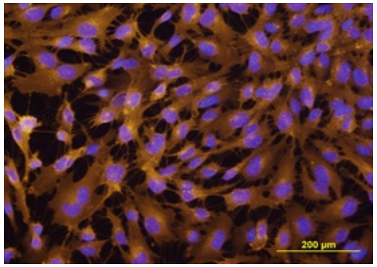
Forrest et al.
Nat Genet 2009

Balwierz et al.
Genome Res 2014

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

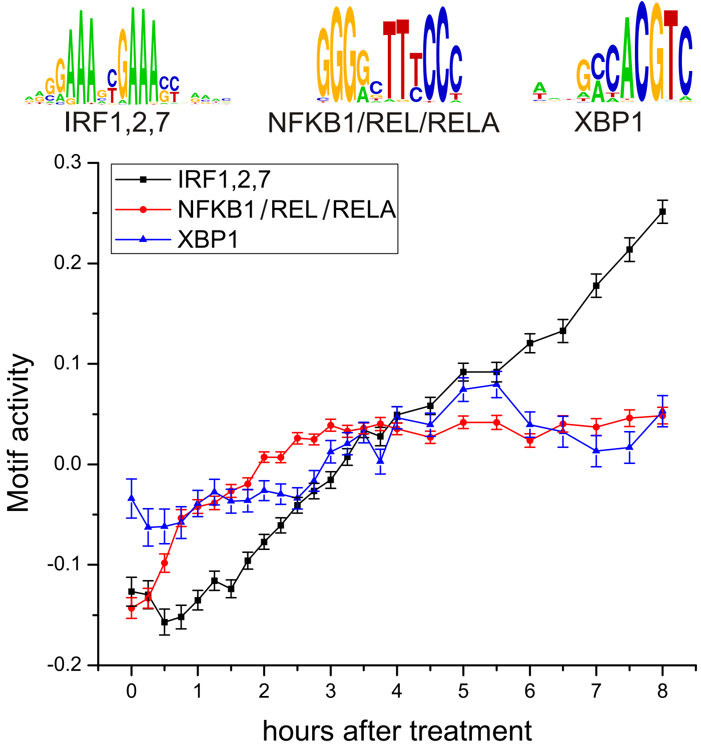
SIB
Swiss Institute of Bioinformatics

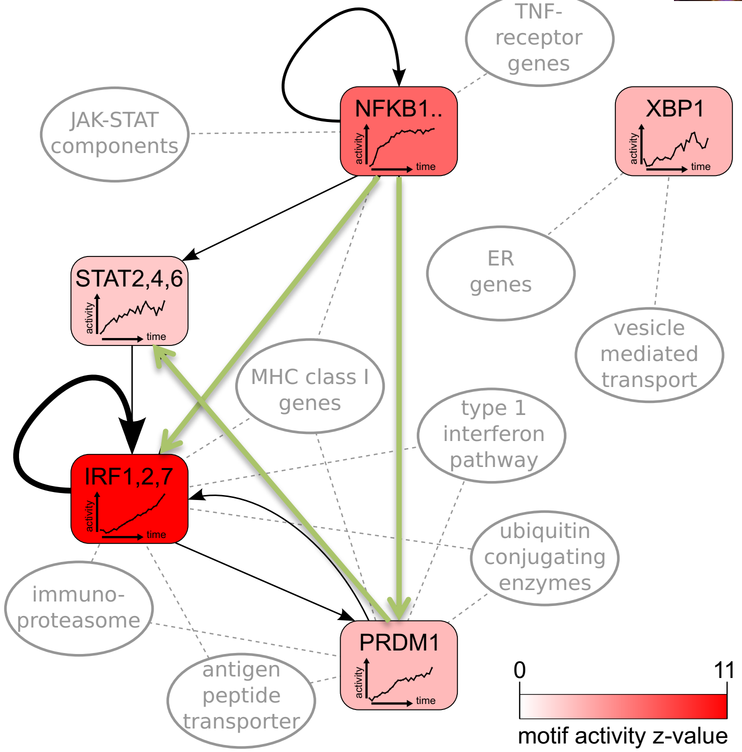
Example: Response of Human umbilical vein endothelial cells to treatment with TNFα



Time course measurements: Wada *et al.* A Wave of nascent transcription on activated human genes. *PNAS* 2009

Top 3 most significant motifs





Predicted regulatory interaction

Predicted interaction with experimental support.

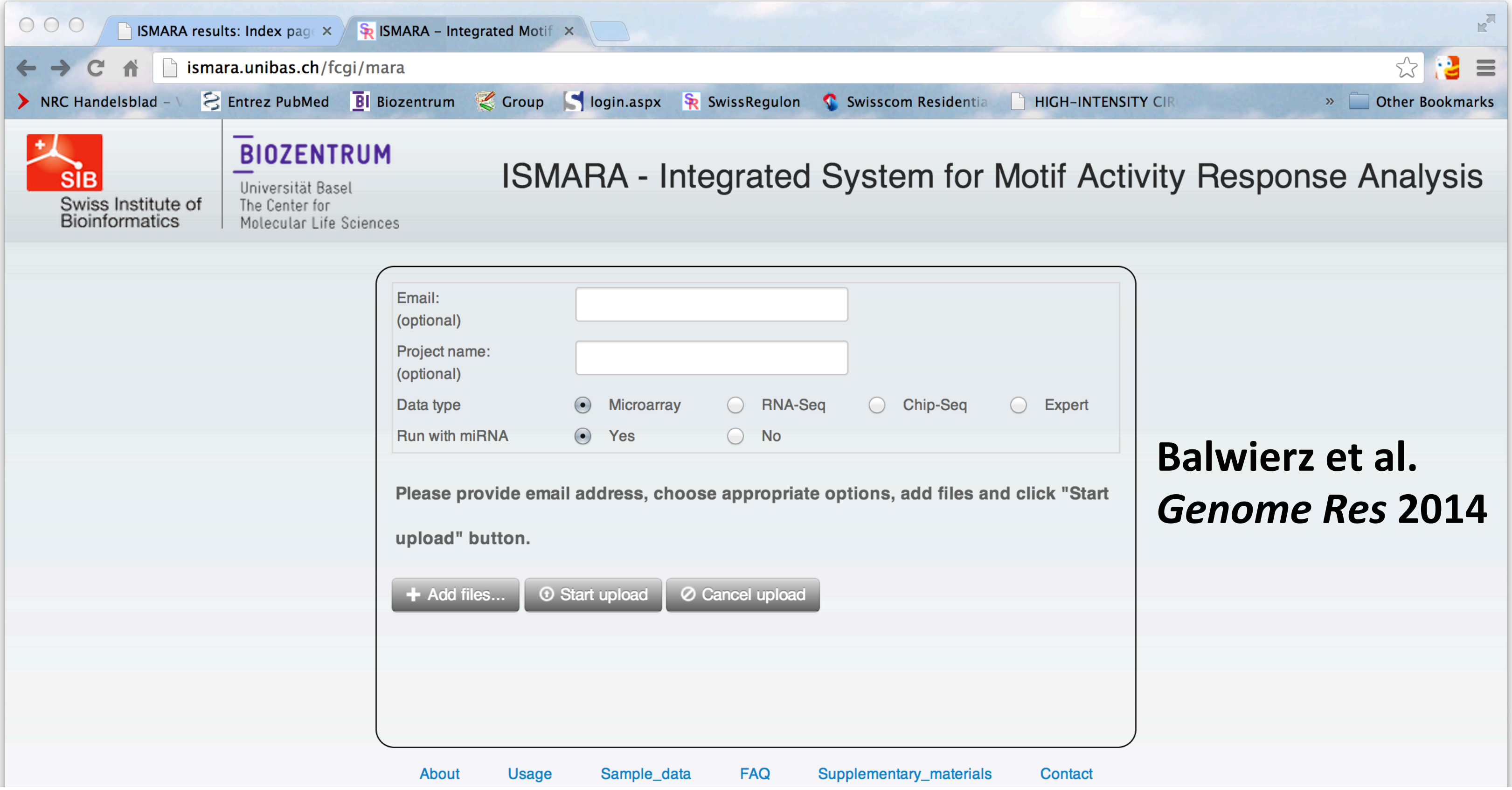
Enriched target gene category

http://ismara.unibas.ch

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Completely automated prediction of regulatory interactions from high-throughput data


ISMARA results: Index page
ISMARA – Integrated Motif
ismara.unibas.ch/fcgi/mara
NRC Handelsblad
Entrez PubMed
Biozentrum
Group
login.aspx
SwissRegulon
Swisscom Residentia
HIGH-INTENSITY CIR
Other Bookmarks
SIB
Swiss Institute of Bioinformatics
BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

ISMARA - Integrated System for Motif Activity Response Analysis

Email:
(optional)

Project name:
(optional)

Data type ● Microarray ○ RNA-Seq ○ Chip-Seq ○ Expert

Run with miRNA ● Yes ○ No

Please provide email address, choose appropriate options, add files and click "Start upload" button.

Balwierz et al.
***Genome Res* 2014**

Upload micro-array, RNA-seq, or ChIP-seq data and predict:

- Key regulators (TFs/miRNAs) in the system.
- Regulator activities across the input samples.
- Sets of target genes and pathways for each regulator.
- The regulatory sites on the genome through which the regulators acts.
- Interactions between the regulators.

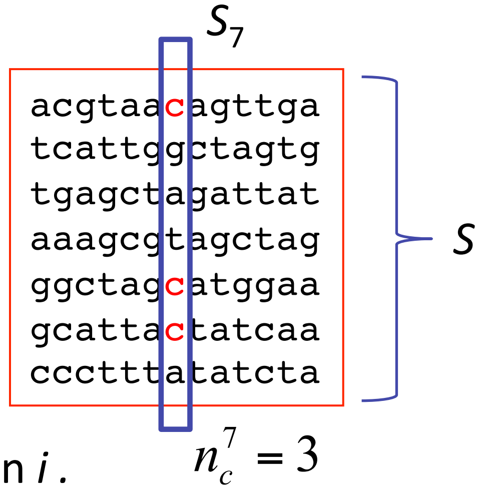
Modeling TF binding specificity

Going beyond position-specific weight matrices

Probability for a set of sequences to derive from a common WM

Probability of observing the set of sequences *S* when sampling from the *known* WM *w*:

$$P(S\,|\,w)=\prod_{i=1}^{l}P(S_i\,|\,w^i)=\prod_{i=1}^{l}\left[\prod_{\alpha}\left(w_\alpha^i\right)^{n_\alpha^i}\right]$$



n_α^i = number of times letter *α* appears at position *i* in *S.*

$w^i=\left(w_a^i,w_c^i,w_g^i,w_t^i\right)$ $\quad w_\alpha^i$ = probability letter *α* appears at position *i* .

- The weight matrix *w* is an *unknown* variable in our model.
- Probability theory prescribes that we should introduce a *prior probability distribution* for it and *integrate it out* of our probability.
- Using the Dirichlet prior:

$$P(w^i)\propto\prod_{\alpha}\left(w_\alpha^i\right)^{\lambda-1}$$

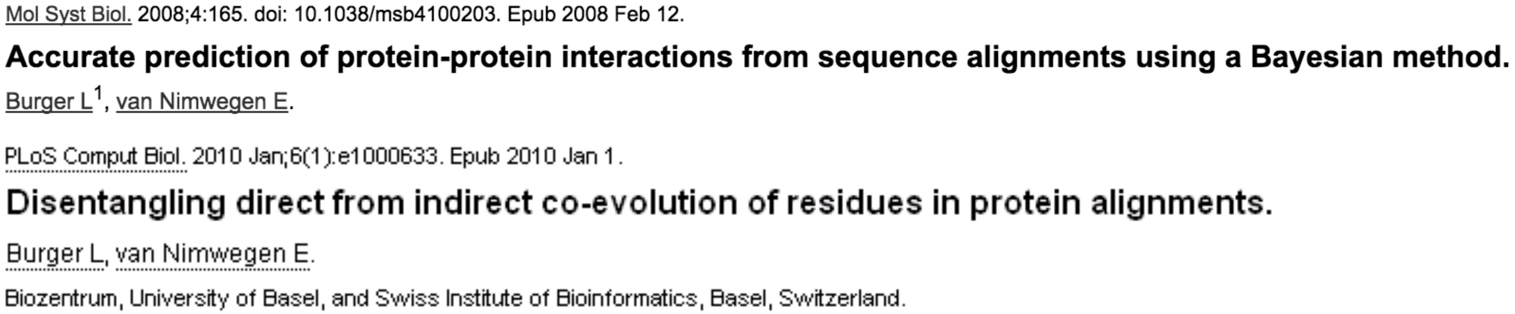
- One obtains:

$$P(S^i)=\int P(S^i\,|\,w^i)P(w^i)\,dw^i=\frac{\Gamma(4\lambda)}{\Gamma(n+4\lambda)}\prod_{\alpha}\frac{\Gamma\left(n_\alpha^i+\lambda\right)}{\Gamma(\lambda)}$$

Including pairwise dependencies

We extend the PWM to a Dinucleotide Weight Tensor (DWT) model that *allows arbitrary pairwise dependencies* between positions.

Lukas Burger



Mol Syst Biol. 2008;4:165. doi: 10.1038/msb4100203. Epub 2008 Feb 12.

Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method.

Burger L[1], van Nimwegen E.

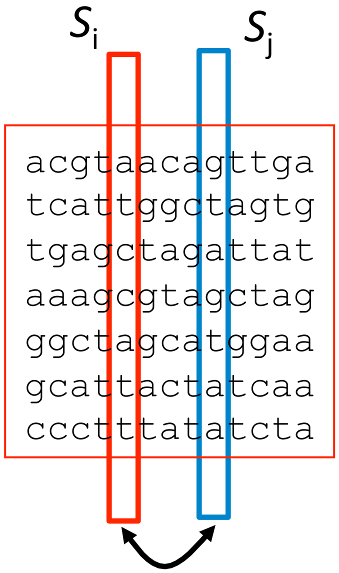
PLoS Comput Biol. 2010 Jan;6(1):e1000633. Epub 2010 Jan 1.

Disentangling direct from indirect co-evolution of residues in protein alignments.

Burger L, van Nimwegen E.

Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland.

Probability for a pair of columns under a DWT



```
acgtaacagttga
tcattggctagtg
tgagctagattat
aaagcgtagctag
ggctagcatggaa
gcattactatcaa
cccttatatcta
```

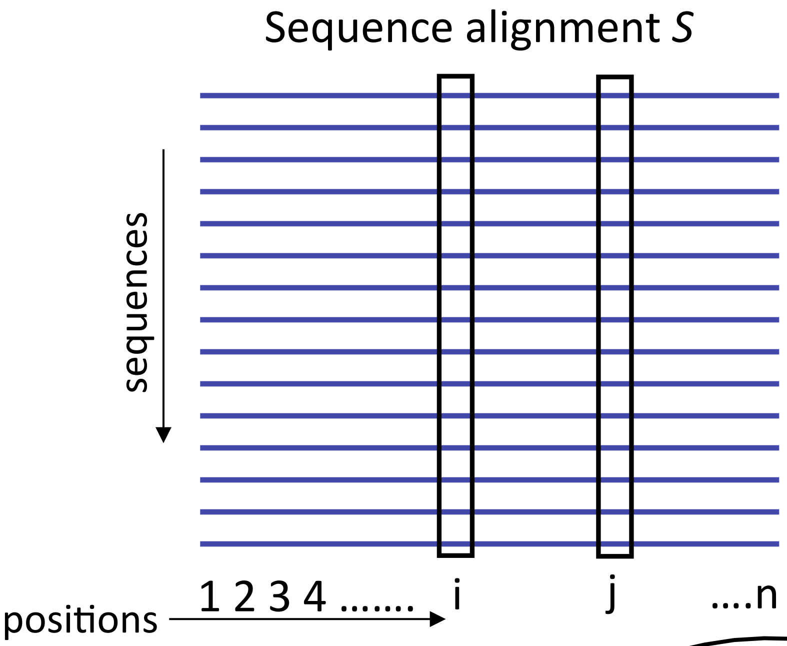
$$S_i = \{n^i_\alpha\} \qquad S_j = \{n^j_\beta\} \qquad (S_i, S_j) = \{n^{ij}_{\alpha\beta}\}$$

$w^{ij}_{\alpha\beta}$ = Probability for the pair of nucleotides *α,β* to occur at positions (*i,j*).

$$P(S_i,S_j) = \int P(S_i,S_j \,|\, w^{ij}) P(w^{ij}) dw^{ij} = \frac{\Gamma(16\tilde{\lambda})}{\Gamma(n+16\tilde{\lambda})} \prod_{\alpha,\beta} \frac{\Gamma(n^{ij}_{\alpha\beta}+\tilde{\lambda})}{\Gamma(\tilde{\lambda})}$$

Likelihood ratio: $R_{ij} = \dfrac{P(S_i,S_j)}{P(S_i)P(S_j)} \approx \exp(nI_{ij})$

Probability given a dependence tree



PWM model:

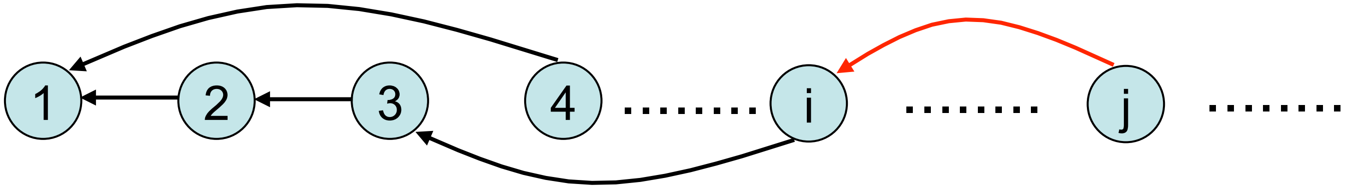
- Each position is independent:

$$P(S) = \prod_i P(S_i)$$

DWT model:

- The probability of observing a given nucleotide at a position *i* of the alignment depends on the nucleotide at *one* other position *π*(*i*).
- The set of `parents' *π*(*i*) of all positions *i* determine a *spanning tree* of the set of positions.

Dependence tree π:



Factorization:

$$P(S|\pi) = P(S_1)P(S_2|S_1)P(S_3|S_2)P(S_4|S_1)\cdots P(S_i|S_3)\cdots P(S_j|S_i)\cdots$$

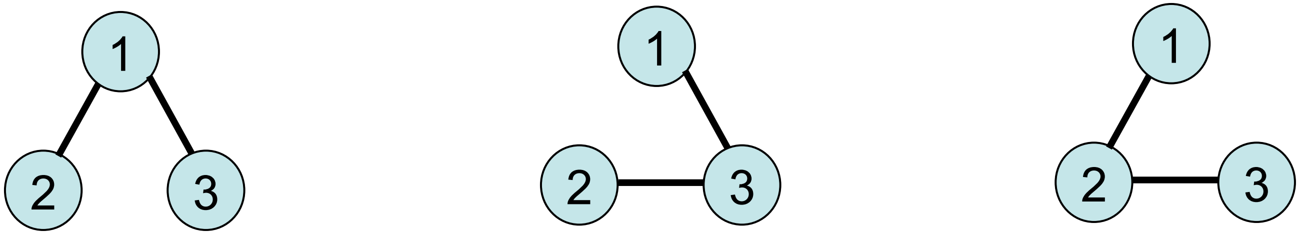
$$P(S|\pi) = P(S_r)\prod_{i\neq r}\frac{P(S_i,S_{\pi(i)})}{P(S_{\pi(i)})} = \prod_i P(S_i)\prod_{(i,j)\in\pi} R_{ij}$$

Summing over spanning trees

Since we do not know the spanning tree, probability theory prescribes we should sum over all possible spanning tree (with uniform prior):

$$P(S)=\sum_{T}\frac{P(S\,|\,\pi)}{|\,\pi\,|}=\frac{1}{|\,\pi\,|}\prod_{i}P(S_i)\sum_{\pi}\left[\prod_{(i,j)\in\pi}R_{ij}\right]$$

Example: for 3 positions we would sum over the three possible spanning trees:



$$P(S)\propto R_{12}R_{13}+R_{13}R_{23}+R_{12}R_{23}$$

Using Kirchhoff/Matrix-tree theorem

Laplacian matrix of *R*: $L(R)_{ij}=\delta_{ij}\sum_{k}R_{ik}-R_{ij}$

Define: *D*(*R*) = Any minor (determinant) of the L(*R*), then: $\sum_{\pi}\left[\prod_{(i,j)\in\pi}R_{ij}\right]=D(R)$

Final probability under the DWT model: $P(S)=\dfrac{D(R)}{|\,\pi\,|}\prod_{i}P(S_i)$

Predicting TFBS and motif finding with DWTs

s

acgtaaag**acgtagcgatcgaa**gagatctcggagcgtaagcagcaacgggatcagagagcaaattat

S

```
acgtaacagttga
tcattggctagtg
tgagctagattat
aaagcgtagctag
ggctagcatggaa
gcattactatcaa
ccctttatatcta
```

Probability that a sequence *s* derives from the *same* motif as the set of sequences *S*:

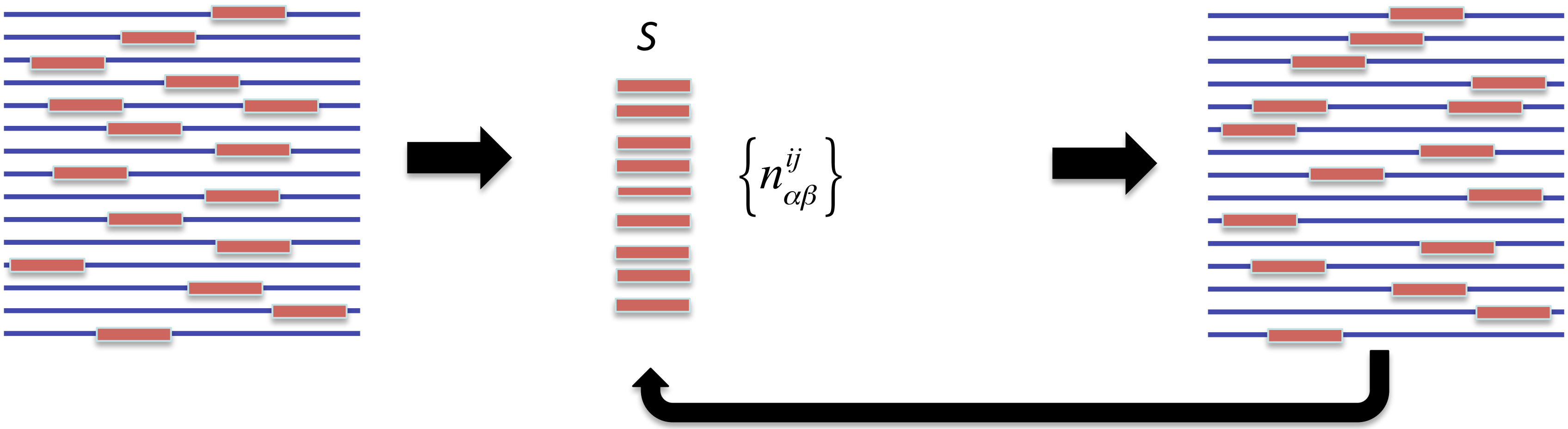
$$P(s\,|\,S) = \frac{P(s,S)}{P(S)} = \frac{D(R(s,S))}{D(R(S))}\prod_i \frac{P(S_i,s_i)}{P(S_i)} = \underbrace{\frac{D(R(s,S))}{D(R(S))}}_{\text{dependencies}} \overbrace{\prod_i \frac{n^i_{s_i}+\lambda}{n+4\lambda}}^{\text{PWM part}}$$

Expectation Maximization procedure for motif finding

1.Predict sites with initial motif

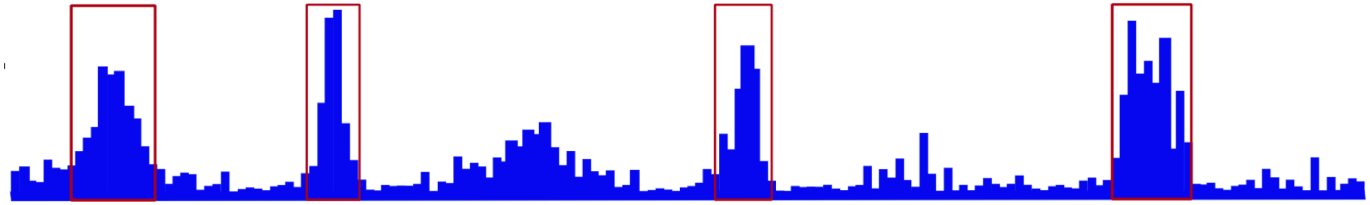
2. DWT defined by dinucleotide counts in sites.

3. Predict sites with current DWT.

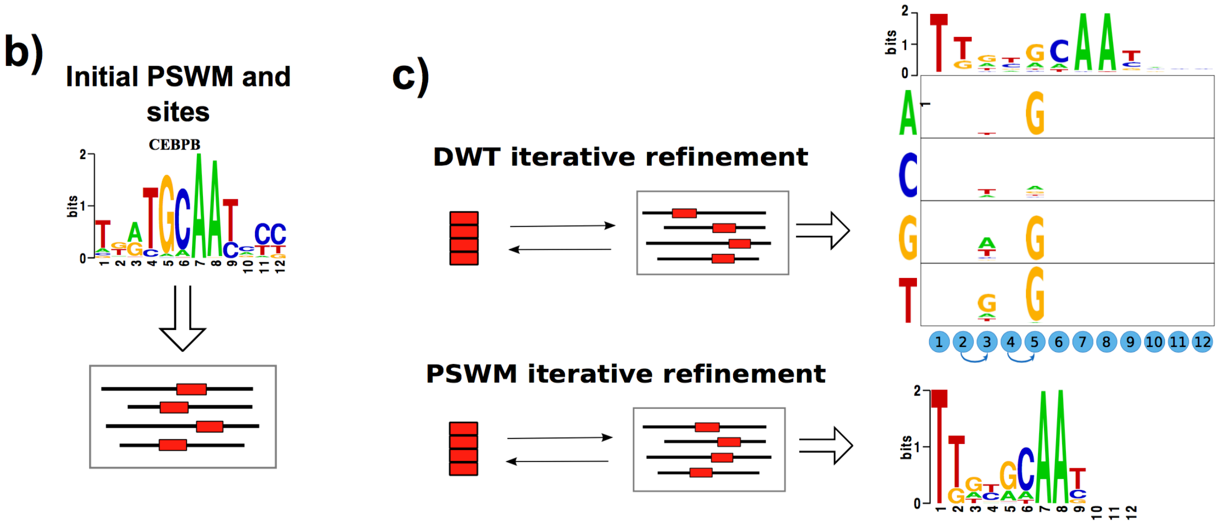


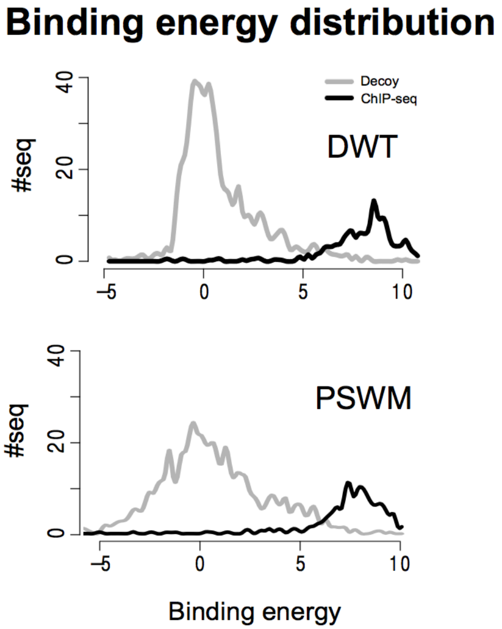
Testing DWT performance on ChIP-seq datasets from ENCODE

- **Data**: ChIP-seq data-sets from ENCODE for 83 different human TFs.
- **Processing of each TF's data-set**:
  - top 1000 peaks from Crunch.
  - Divide into 500 *training regions*, and 500 *test regions*.



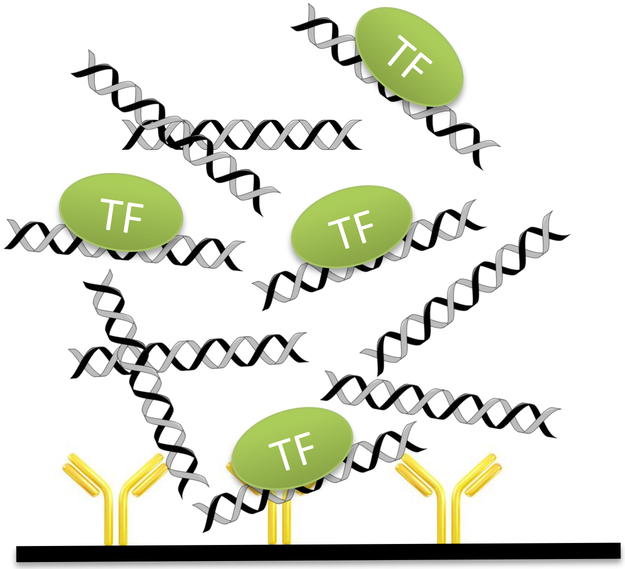
- Fit both a PWM and DWT on the training regions.



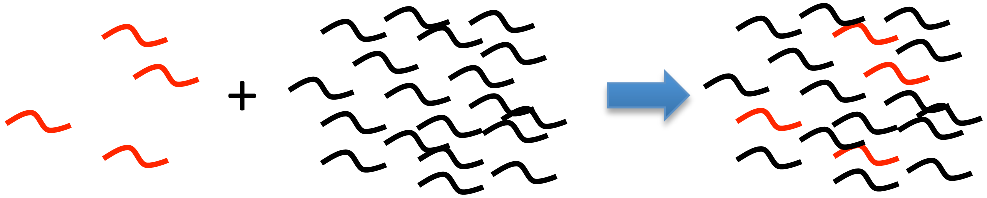


- Calculate enrichment score on the test set mixed with background regions of equal dinucleotide composition.

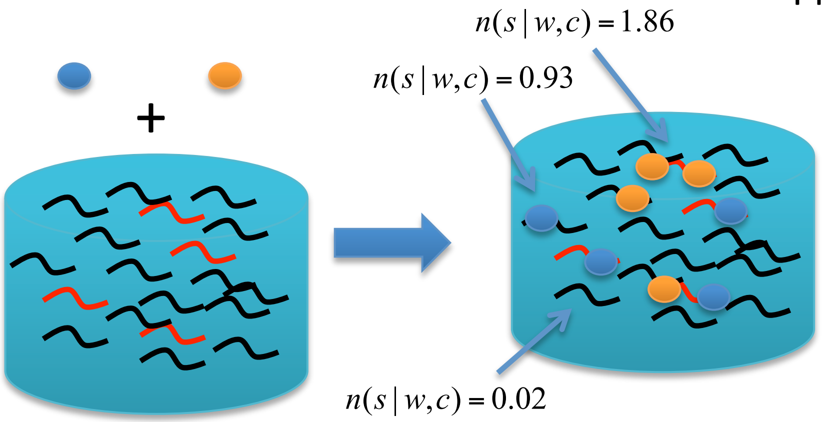
An enrichment score for ChIP-seq



- **Data**: IP ‘fished’ our peak sequences from a much larger collection of DNA fragments.
- **Assumption**: The probability to fish (=IP) a sequence is proportional to the *number of copies of the TF(s)* bound to it.
- **Likelihood model:**
  - Peak sequences *P* + Background sequences ***B*** (= random seqs with same lengths and nucleotide composition).



  - Given a set of motifs *w*, and their concentrations *c*, calculate the expected number of bound TFs $n(s|w,c)$ at each sequence *s*.
  - Probability to IP sequence *s*: $P(s|w,c) = \dfrac{n(s|w,c)}{\sum_{s' \in P \cup B} n(s'|w,c)}$



Probability to IP *all* sequences in *P* and *only* the sequences in *P*: $P(D|w,c) = \prod_{s \in P} P(s|w,c)$

Likelihood for a motif set *w*: $P(D|w) = \max_{c} P(D|w,c)$

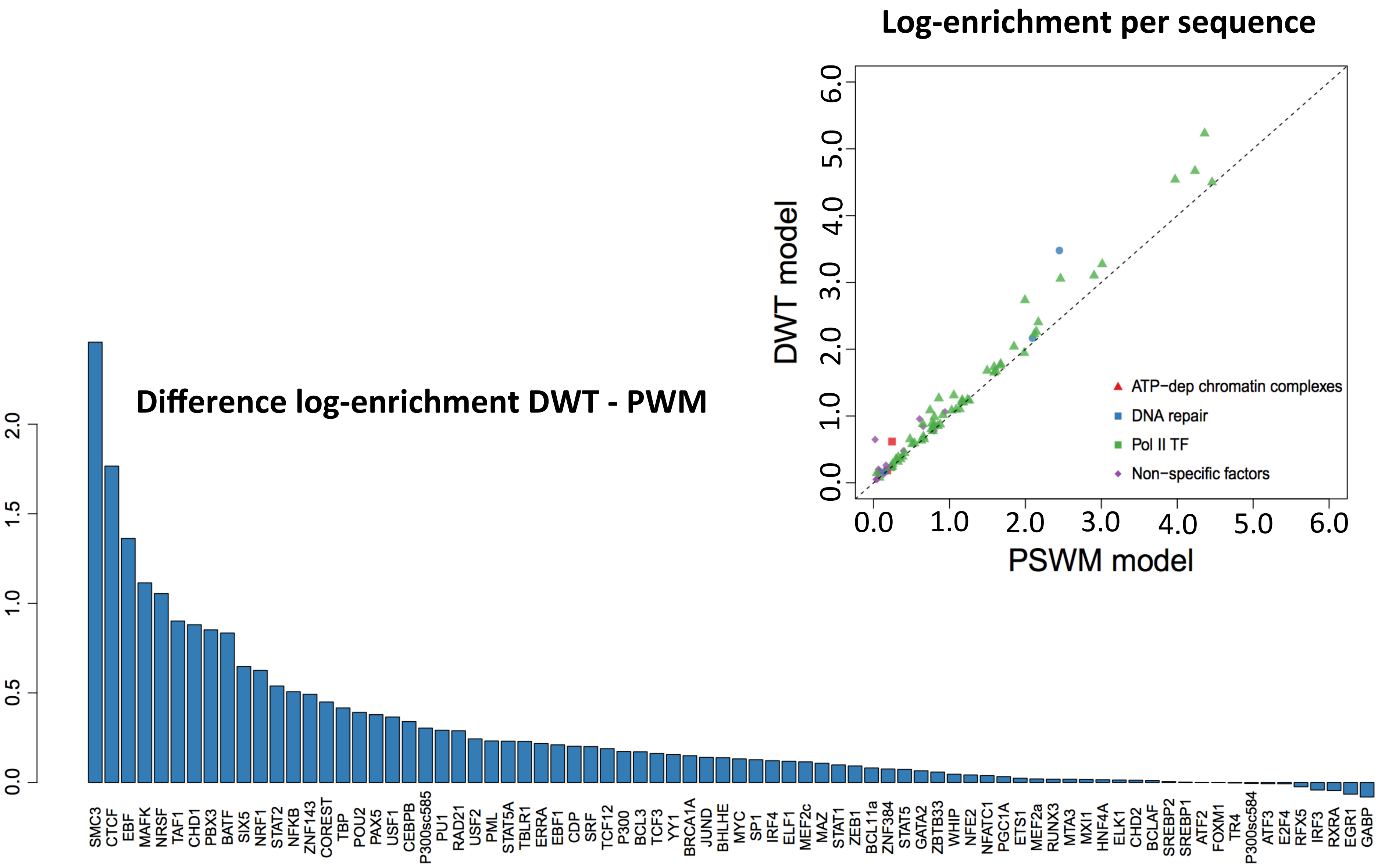
BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

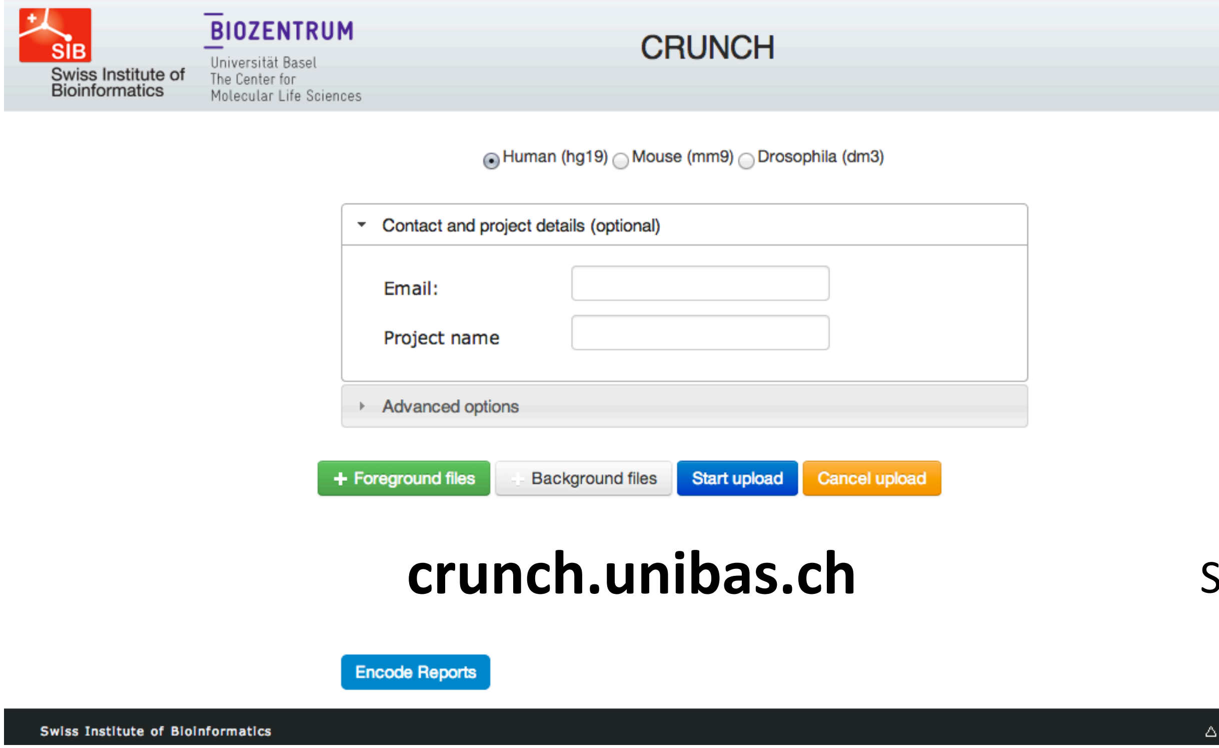
DWTs often outperform PWMs and never overfit



Log-enrichment per sequence

Difference log-enrichment DWT - PWM

CRUNCH: A completely automated webserver for ChIP-seq data analysis



SIB
Swiss Institute of Bioinformatics
BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences
CRUNCH

● Human (hg19) ○ Mouse (mm9) ○ Drosophila (dm3)

Contact and project details (optional)

Email:

Project name

crunch.unibas.ch

Severin Berger

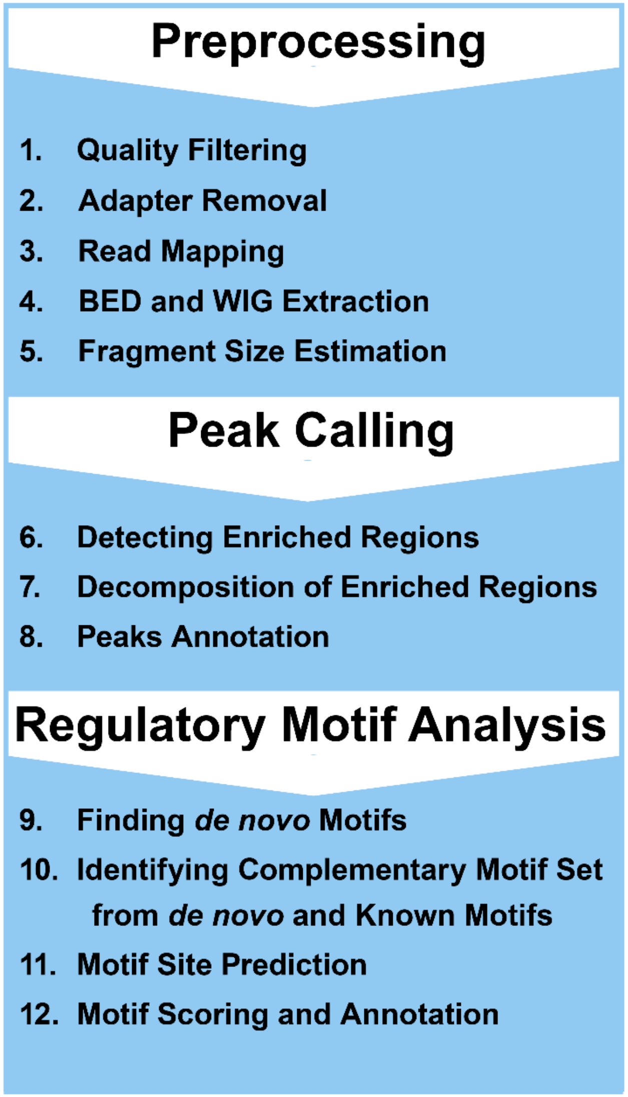
Motivation

- For tools like MARA we would like to automatically process available ChIP-seq data to curate new motifs and annotate where they bind.
- However, ChIP-seq data analysis is still *wild-west*:
  - Almost no standardized procedures even within consortia like ENCODE.
  - Cannot meaningfully compare results from different studies.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Overview of CRUNCH analysis steps



Preprocessing

1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

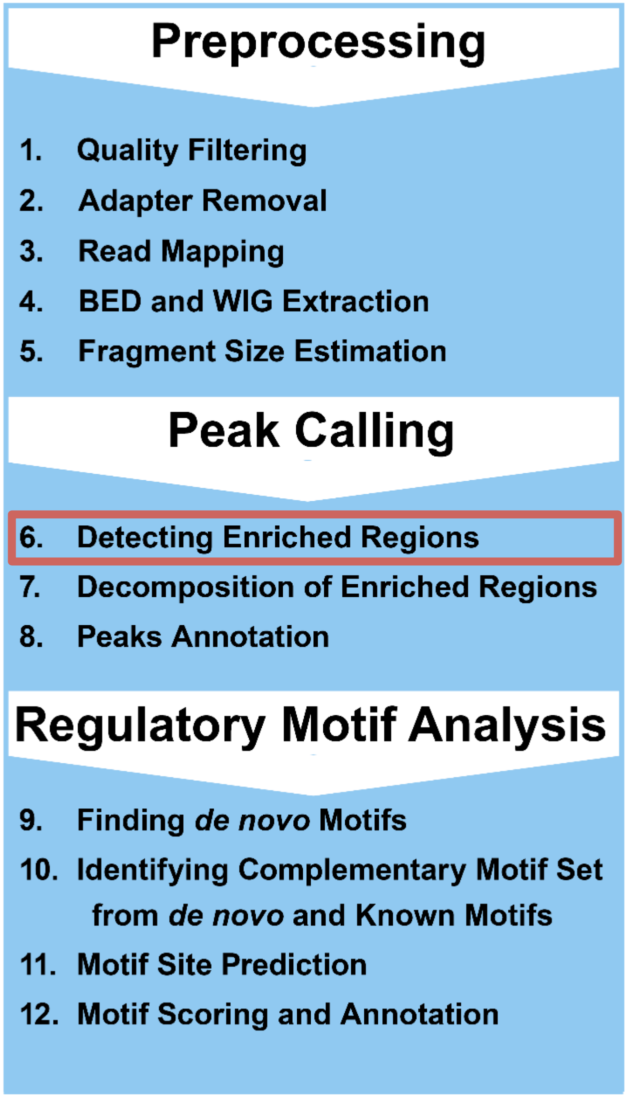
Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

Detecting enriched regions



Preprocessing

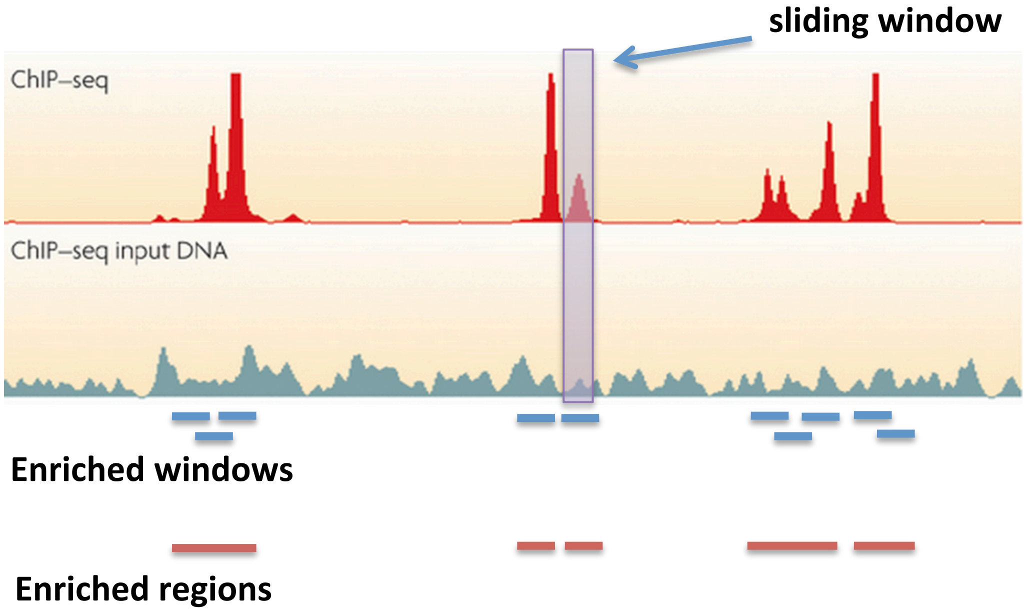
1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

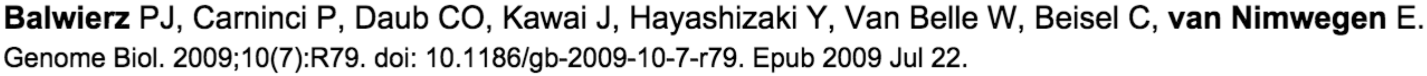


- Slide 500 bp window across genome.
- Quantify significance of the enrichment of ChIP-seq over input DNA.

Bayesian model for identifying enriched regions

Noise model for read-counts in un-enriched windows

- *Multiplicative* noise plus *Poisson* sampling, i.e. as previously developed in:



Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, **van Nimwegen** E.
Genome Biol. 2009;10(7):R79. doi: 10.1186/gb-2009-10-7-r79. Epub 2009 Jul 22.

Variables:

- *n*,*m* = reads in ChIP/input sample.
- *N*,*M* = total reads in ChIP/input sample.
- *σ* = standard-deviation of the multiplicative noise.
- *μ* = shift in average log read-density.

Enrichment *x*:

$$x = \log\left[\frac{n}{N}\right] - \log\left[\frac{m}{M}\right]$$

Probability of observing *x*: $P(x \mid \mu,\sigma) \propto \exp\left[-\frac{(x-\mu)^2}{2\left(2\sigma^2 + \frac{1}{n} + \frac{1}{m}\right)}\right]$

Mixture model

- The enrichment x_i for each window *i* derives from either the noise model or a uniform distribution (= ‘something else’):

$$P(D \mid \mu,\sigma,\rho) = \prod_i \left[P(x_i \mid \mu,\sigma)\rho + \frac{1-\rho}{x_{\max} - x_{\min}}\right]$$

- We fit *μ*, *σ*, and *ρ* to *maximize* $P(D \mid \mu,\sigma,\rho)$, and calculate an enrichment z-score for each window.

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

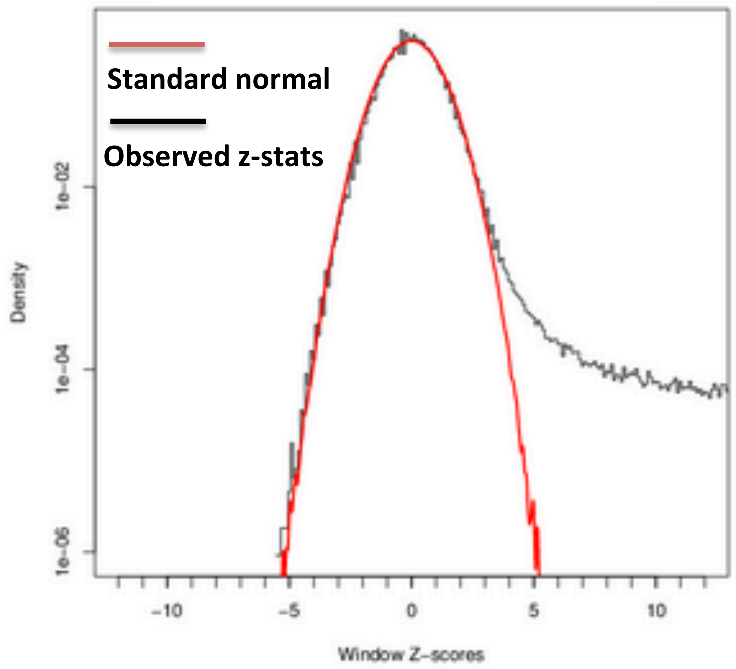
SIB
Swiss Institute of Bioinformatics

The noise model accurately captures the observed genome-wide enrichment statistics

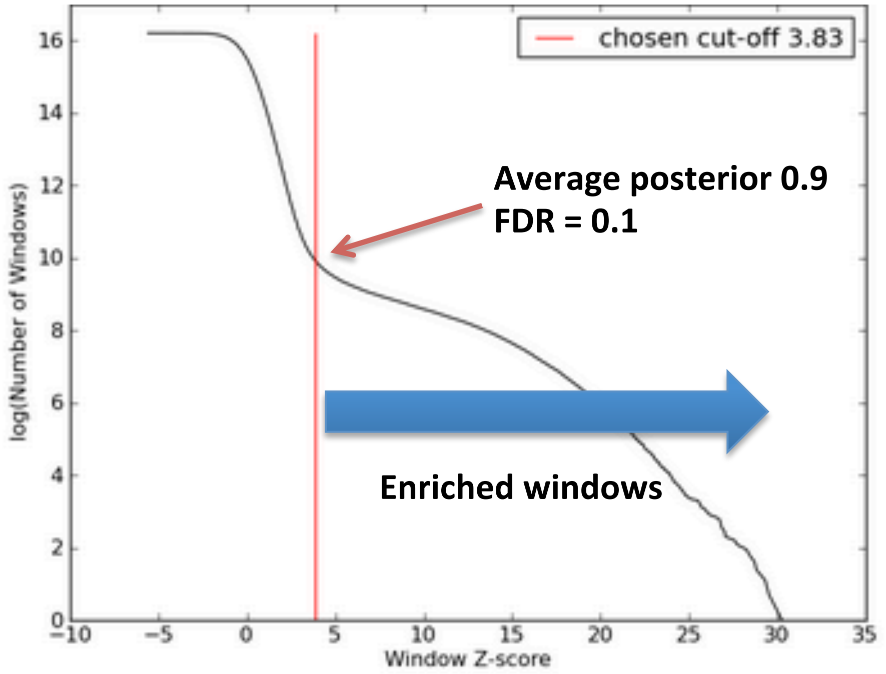
Z-statistic for each window:

$$z_i = \frac{\log\left[\frac{n_i}{N}\right] - \log\left[\frac{m_i}{M}\right] - \mu}{\sqrt{2\sigma^2 + \frac{1}{n_i} + \frac{1}{m_i}}}$$

Distribution of z-scores



Reverse cumulative distribution of z-scores

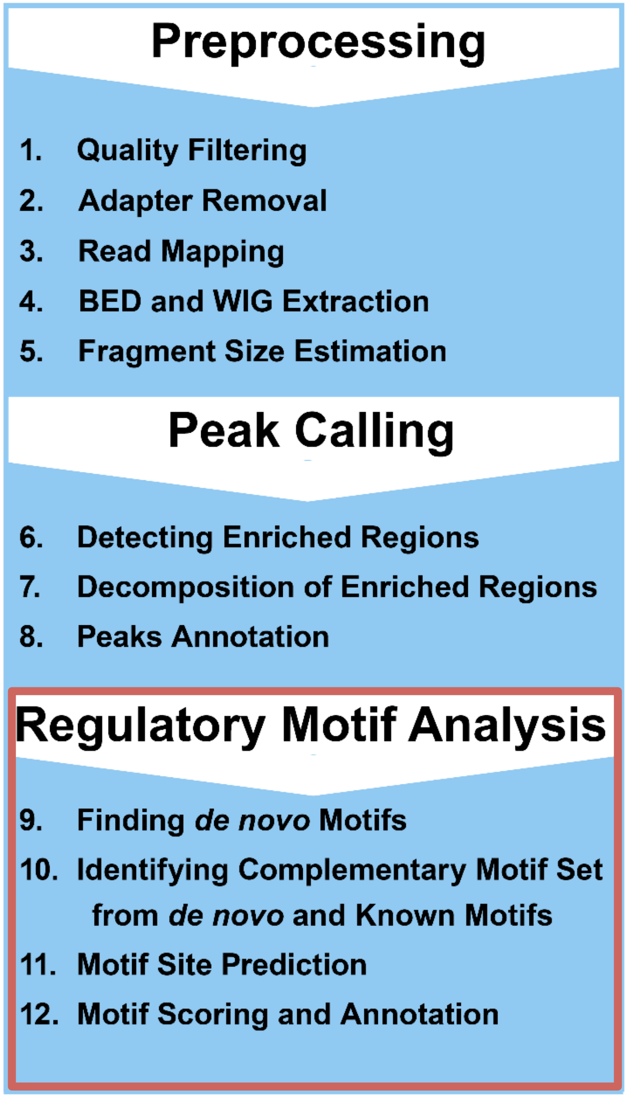


Average posterior 0.9
FDR = 0.1

Enriched windows

As far as we are aware, **ours is the only peak-finder that demonstrably matches the data's statistics**.

Overview of the analysis steps



Preprocessing

1. Quality Filtering
2. Adapter Removal
3. Read Mapping
4. BED and WIG Extraction
5. Fragment Size Estimation

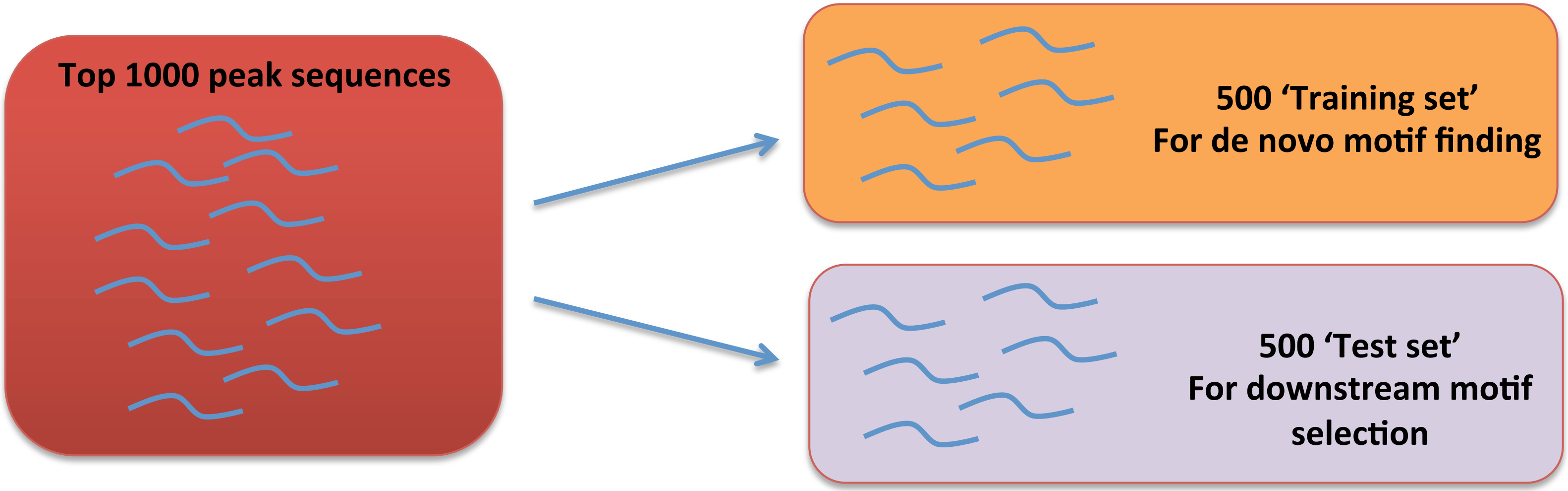
Peak Calling

6. Detecting Enriched Regions
7. Decomposition of Enriched Regions
8. Peaks Annotation

Regulatory Motif Analysis

9. Finding *de novo* Motifs
10. Identifying Complementary Motif Set from *de novo* and Known Motifs
11. Motif Site Prediction
12. Motif Scoring and Annotation

De novo motif finding

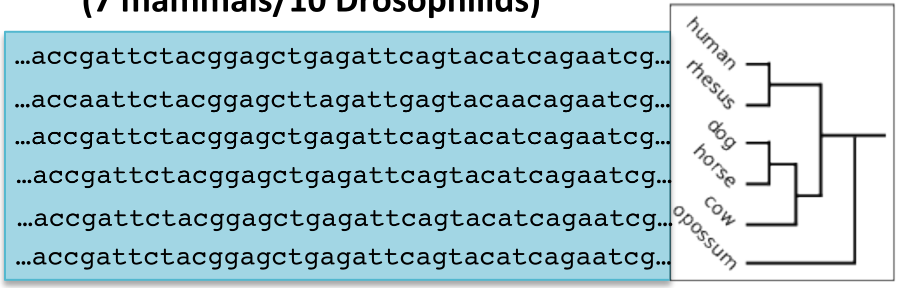


Top 1000 peak sequences

500 'Training set'
For de novo motif finding

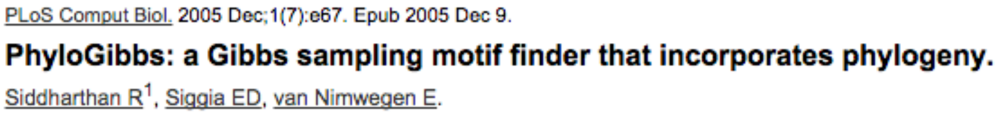
500 'Test set'
For downstream motif selection

1. Align with orthologous regions (7 mammals/10 Drosophilids)



```
…accgattctacggagctgagattcagtacatcagaatcg…
…accaattctacggagcttagattgagtacaacagaatcg…
…accgattctacggagctgagattcagtacatcagaatcg…
…accgattctacggagctgagattcagtacatcagaatcg…
…accgattctacggagctgagattcagtacatcagaatcg…
…accgattctacggagctgagattcagtacatcagaatcg…
```

2. Identify motifs with PhyloGibbs



PLoS Comput Biol. 2005 Dec;1(7):e67. Epub 2005 Dec 9.

PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.

Siddharthan R[1], Siggia ED, van Nimwegen E.

3. Refine motifs with MotEvo

Bioinformatics. 2012 Feb 15;28(4):487-94. doi: 10.1093/bioinformatics/btr695.

MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences.

Arnold P[1], Erb I, Pachkov M, Molina N, van Nimwegen E.

4. Result

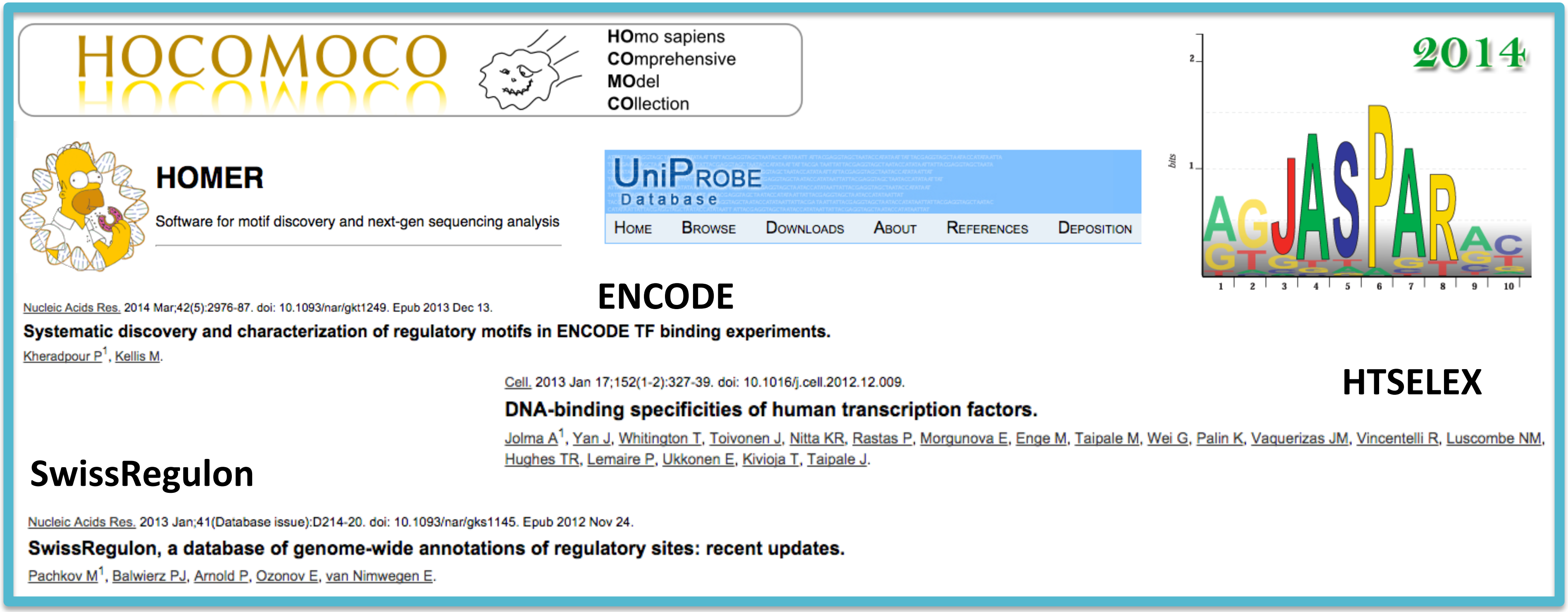
Up to 24 candidate *de novo* motifs

BIOZENTRUM
Universität Basel
The Center for Molecular Life Sciences

SIB
Swiss Institute of Bioinformatics

Library of known motifs

Library of 2325 known motifs (position-specific weight matrices) from:



HOCOMOCO

HOmo sapiens
COmprehensive
MOdel
COllection

HOMER

Software for motif discovery and next-gen sequencing analysis

UniPROBE Database

ENCODE

Nucleic Acids Res. 2014 Mar;42(5):2976-87. doi: 10.1093/nar/gkt1249. Epub 2013 Dec 13.

Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments.

Kheradpour P[1], Kellis M.

HTSELEX

Cell. 2013 Jan 17;152(1-2):327-39. doi: 10.1016/j.cell.2012.12.009.

DNA-binding specificities of human transcription factors.

Jolma A[1], Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J.

SwissRegulon

Nucleic Acids Res. 2013 Jan;41(Database issue):D214-20. doi: 10.1093/nar/gks1145. Epub 2012 Nov 24.

SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates.

Pachkov M[1], Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E.

Task

Find a set of complementary known/*de novo* motifs
that jointly explain the observed binding peaks of the test set.

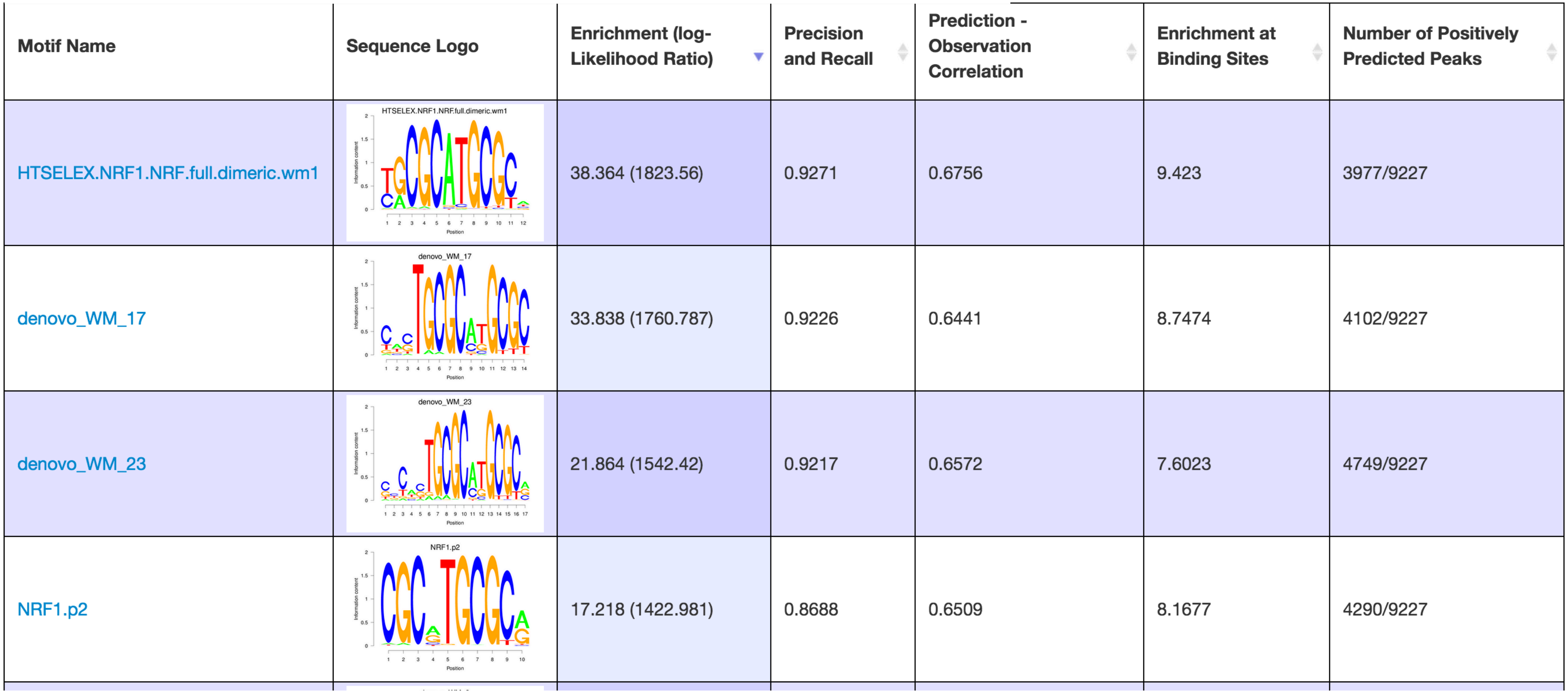
Sorted list of most enriched motifs

Final enrichment score : Per sequence likelihood ratio relative to *randomly selecting sequences:*

$$E_w = \left[\frac{P(D\,|\,w,c)}{P(D\,|\,\text{random})}\right]^{1/|P|} = \left[\prod_{s\in P}\frac{n(s\,|\,w,c)}{\langle n\rangle_B}\right]^{1/|P|}$$

$|P| =$ Number of binding peaks.

We sort all known and *de novo* motifs by their enrichment.
Example (NRF1 ChIP-seq):



Motif Name	Sequence Logo	Enrichment (log-Likelihood Ratio)	Precision and Recall	Prediction - Observation Correlation	Enrichment at Binding Sites	Number of Positively Predicted Peaks
HTSELEX.NRF1.NRF.full.dimeric.wm1	HTSELEX.NRF1.NRF.full.dimeric.wm1	38.364 (1823.56)	0.9271	0.6756	9.423	3977/9227
denovo_WM_17	denovo_WM_17	33.838 (1760.787)	0.9226	0.6441	8.7474	4102/9227
denovo_WM_23	denovo_WM_23	21.864 (1542.42)	0.9217	0.6572	7.6023	4749/9227
NRF1.p2	NRF1.p2	17.218 (1422.981)	0.8688	0.6509	8.1677	4290/9227

Selecting an optimal set of complementary motifs

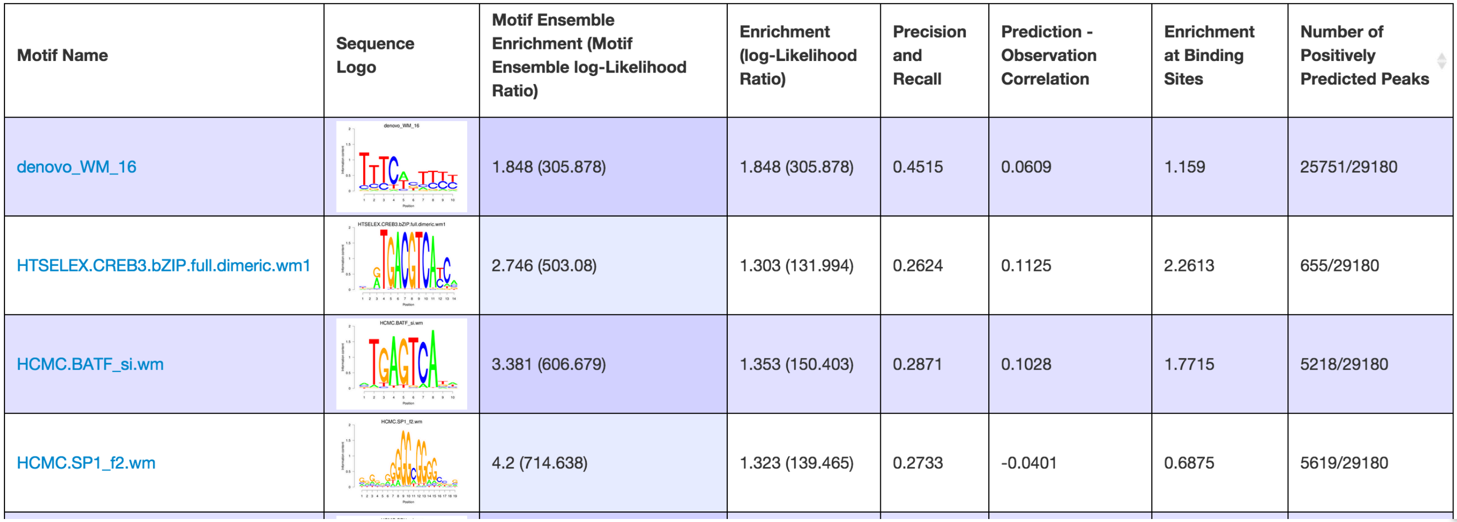
Initialize motif set {w} with best motif w.

Iterate:

1. For each of the remaining motifs w', add w' to {w}, and calculate new $E_{\{w\}}$.
2. Select w' that maximizes $E_{\{w\}}$ and add to the set {w}.

Stop when the enrichment increases by less than 5%.

Example: ATF2 from ENCODE

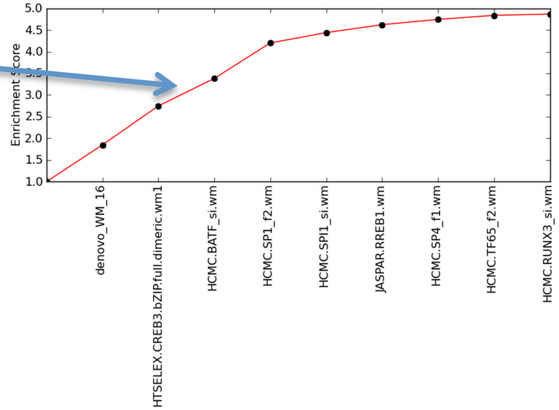


Motif Name	Sequence Logo	Motif Ensemble Enrichment (Motif Ensemble log-Likelihood Ratio)	Enrichment (log-Likelihood Ratio)	Precision and Recall	Prediction - Observation Correlation	Enrichment at Binding Sites	Number of Positively Predicted Peaks
denovo_WM_16		1.848 (305.878)	1.848 (305.878)	0.4515	0.0609	1.159	25751/29180
HTSELEX.CREB3.bZIP.full.dimeric.wm1		2.746 (503.08)	1.303 (131.994)	0.2624	0.1125	2.2613	655/29180
HCMC.BATF_si.wm		3.381 (606.679)	1.353 (150.403)	0.2871	0.1028	1.7715	5218/29180
HCMC.SP1_f2.wm		4.2 (714.638)	1.323 (139.465)	0.2733	-0.0401	0.6875	5619/29180

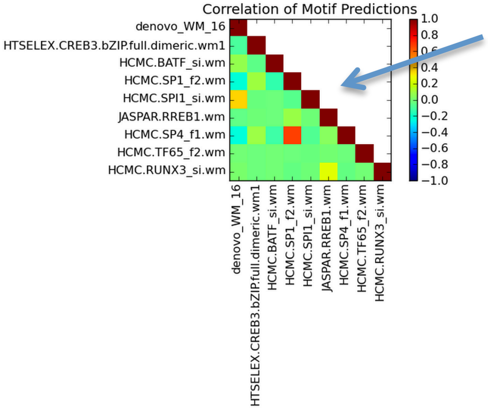
Occurring in most peaks but not specific.

More specific secondary motifs.

Contribution and Correlation Plots hide



Motif combination better explains the binding data.

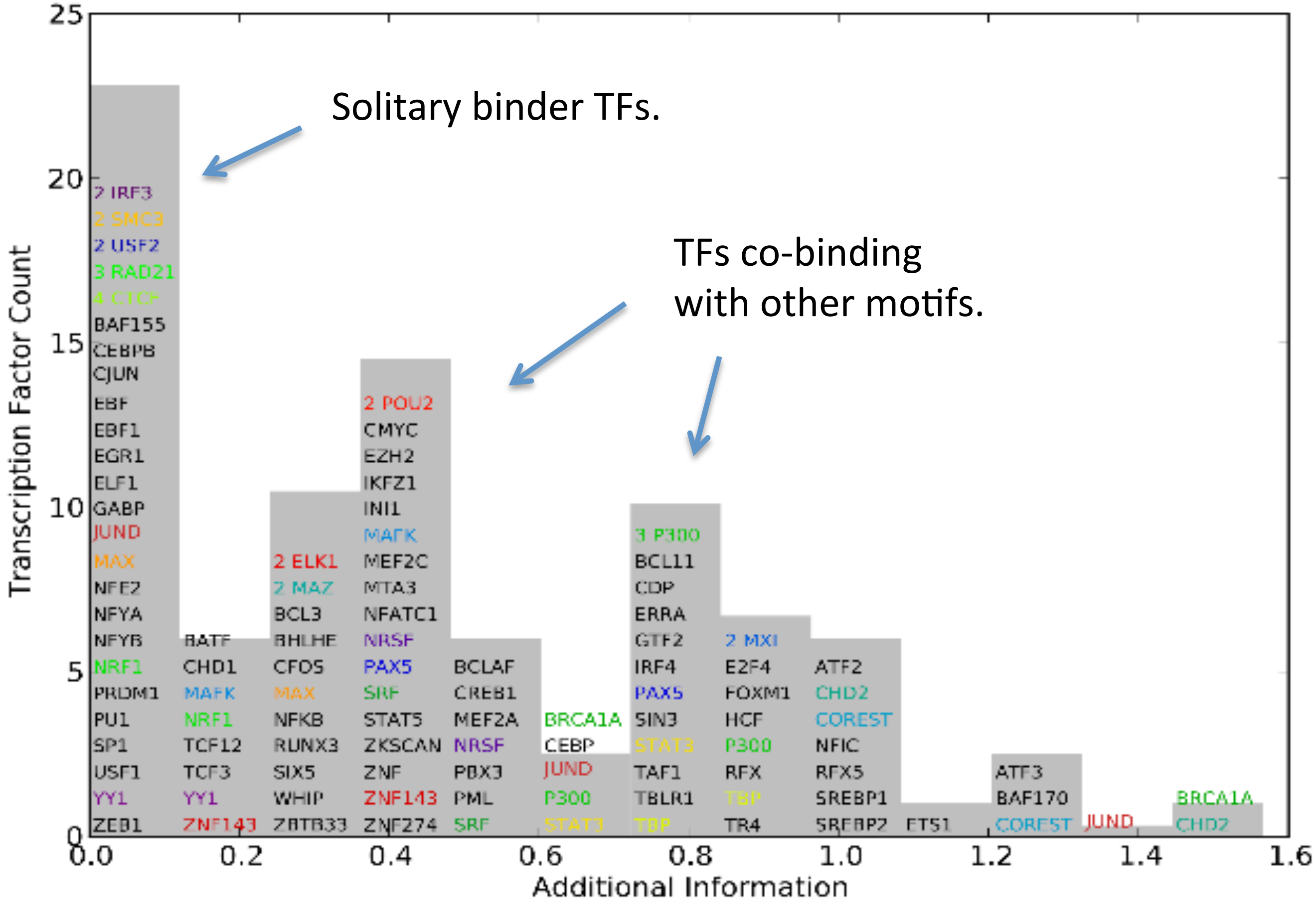


Co-occurrence of sites for different motifs.

We observe two types of TFs:

Solitary binders vs. TFs co-binding with other TFs

Histogram: information in secondary motifs



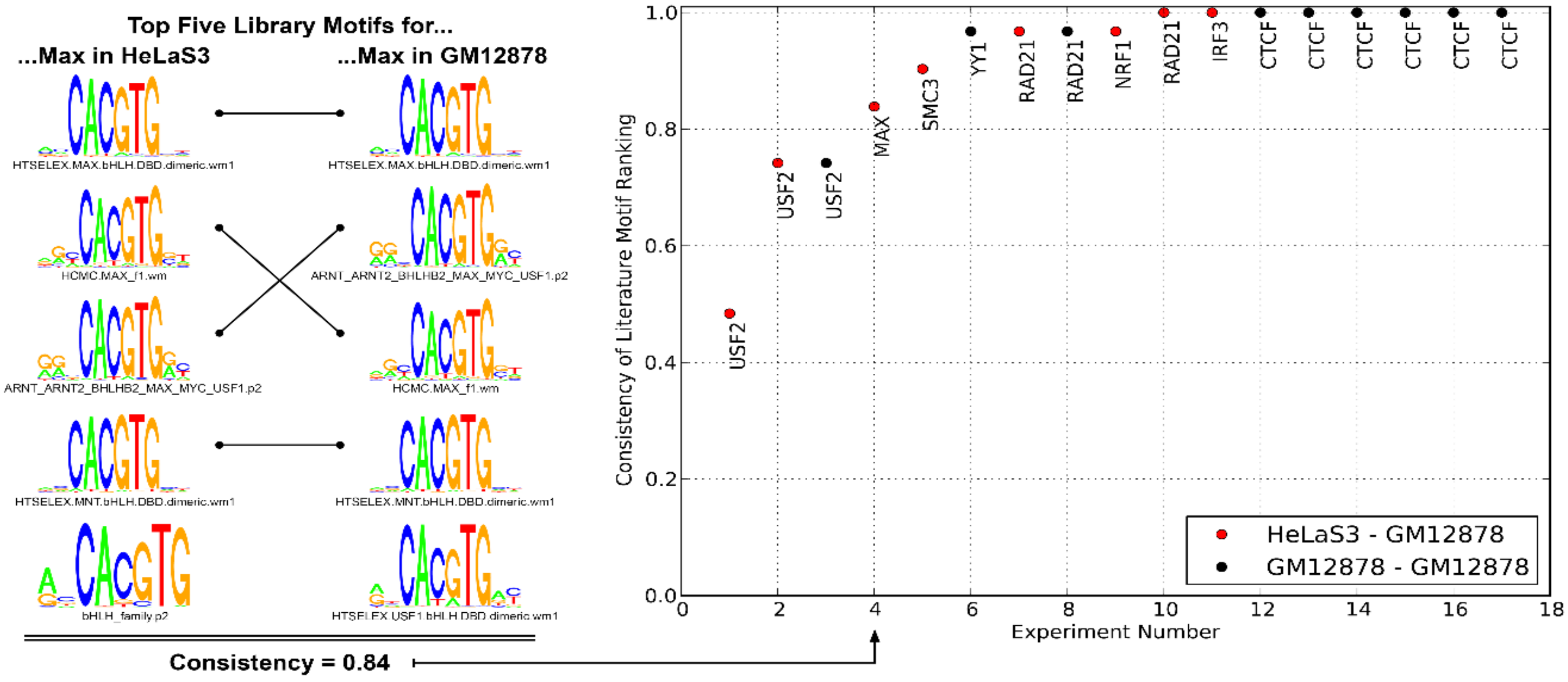
Solitary binder TFs.

TFs co-binding with other motifs.

Top motifs for a TF are consistent across experiments

- Top enriched motifs for a TF are highly consistent across different cell lines/experiments.
- Even when motifs are extremely similar!

Consistency in top motifs across cell lines



This suggests we can select a 'best' motif for each solitary TF in a meaningful way.

Summary and acknowledgments

Crunch:

- Automated webserver for comprehensive ChIP-seq analysis.
- Realistic statistical model.
- Explain the binding peaks in terms of a complementary set of motifs.

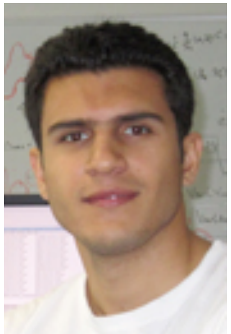
Check BioaRxiv in the coming days for the papers!

Dinucleotide Weight Tensors:

- Rigorous Bayesian model allowing arbitrary dependencies.
- Zero tunable parameters.
- DWTs never overfit and outperform PWMs for many TFs.
- Source code for motif finding and TFBS prediction using DWTs.

Severin Berger
CRUNCH

Lukas Burger
Original DWT model



Saeed Omidi
DWTs for TFBS prediction